Analysis of Patient Data

By Ebenezer Omari

1. Descriptive Introduction:

This report analyzes the data of large number of patients (approximately 2000). The data contains some demographic and simple measurements (height, weight, etc). The data also second contains the results of 5 medical tests where the results range from 0 to 100. It also has an indicator of whether or not the patient has a disease (0=yes, 1=no)

To prepare the data for analysis, the following steps were taken:

1. Merging of the two data using dplyr in R
2. Checked for duplicates
3. Checked any rows with missing values
4. Checked for Na's
5. Replaced all Na's in all variables with their median value
6. Checked for Na's for verification
7. Converted the disease variable to a factor before visualization
8. Checked the structure of the dataset

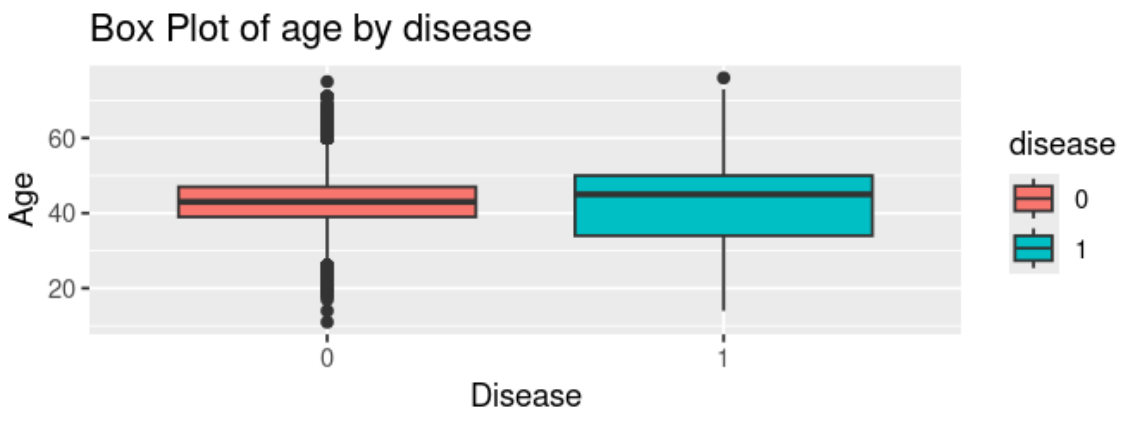
2. Exploratory Data Analysis

Summary of Age in the Patient Data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11.00	37.00	44.00	42.84	49.00	76.00

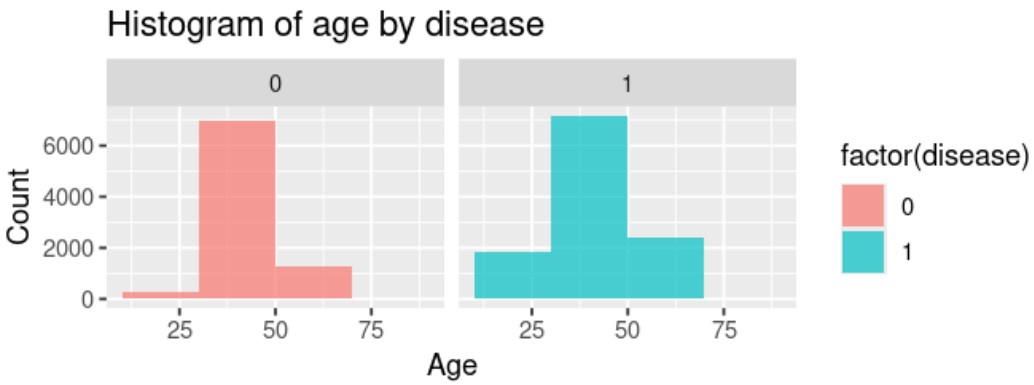
The summary of the age variable gives insight to the distribution of ages in dataset:
The age ranges from 11 years to 76 years.
The average is 43 years.
Half of the participants are younger than 44 years (median) and half are older.
25% of the participants are younger than 37 years (1st quartile).
25% of the participants are older than 49 years (3rd quartile).
The mean value is less than the median, indicating the data is negatively skewed

A Data Visualization of Age by Disease Using Box Plot



Analysis from the Box Plot
The Box plot compared the age for two groups: those without the disease(0) and
those with the disease(1)
The median age for those with the disease appears to be slightly higher than for those without the disease
The box for those without the disease is smaller indicating less variability in age.
The box for those with the disease is larger suggesting more age variability among those with the disease
Both groups have outliers, represented by individual points on beyond the whiskers
The group with no disease has more visible outliers
The age range for those without the disease seems to be more concentrated.
Those with the disease show a wider spread of ages, with a higher median and and more variability
The group with the disease appears slightly skewed positively and the group without the disease is evenly distributed.
In summary while there are some differences in age between those with and without the disease or while age may play a role, the presence of outliers suggest age alone may not be a definite predictor of disease status

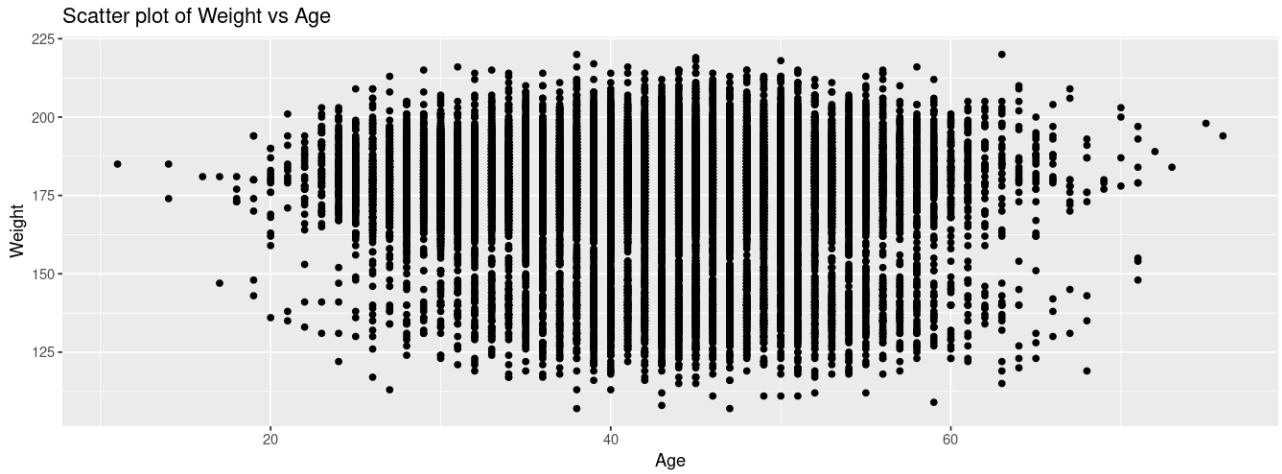
A Data Visualization of Age by Disease Using Histogram



Analysis from the graph
The plot is divided into two panels, one for each disease(0 and 1)
The x-axis represent the age, ranging from 0 to 100 years
The y-axis shows the count of individuals in each age group
Disease = 0(left panel; pink):
The absence of no disease distribution appears to be skewed. Also there is a very high peak in the 25 - 50 age range. A sharp decline occurs after age 50. Very few individuals are the oldest age group(50 - 75)
Disease = 1(right panel; blue):
There's a noticeable peak in the 25-50 age range, but it's less pronounced than in the disease=0 group. There's a more gradual decline in the 50-75 age range.
In summary, those with the disease group has a more even distribution across age groups. Those without the disease is heavily concentrated in the younger age range.

Younger individuals (25-50) without the disease are over represented in the sample. The presence of the disease seems to be associated with a more even age distribution, possibly suggesting it affects people across a broader age spectrum.
The disease=0 group's sharp decline after 50 could indicate that older individuals without the disease are underrepresented in the sample, or that the disease becomes more common with age.
There might be potential biases; The sample might not be representative of the general population, given the strong skew in the absence of disease group. There could be age-related factors influencing disease diagnosis or study participation.

A Data Visualization of Age by Disease Using Scatterplot



Analysis from the scatter plot
The age ranges from 15 to 75 years. Weight ranges from 125 to 225 kgs. From the Plot there is a dense concentration of data points, indicating a large sample size. The data forms a roughly oval shape widening as the age increases. The appears to be a slightly positive correlation between age and weight. As the age increases it appears the upper limit tends to increase more noticeably than the lower limit. As evidenced by the widening spread of points, weight variability increases age. There is noticeably a dense cluster of points in the middle age range (30 – 50years) A few individuals have weights above 200 kgs across different age groups. The increasing variability in weight with age suggests that factors influencing weight become more diverse or impactful as people get older.

3. Further Analysis using confirmatory Data Analysis

A. Logistic Regression
Ho = Age is not a factor to contracting the disease
H1 = Age is a factor to contract the disease

Analysis:
Call:
glm(formula = disease ~ age, family = "binomial", data = project)

Deviance Residuals:
Min 1Q Median 3Q Max

-1.446 -1.300   1.004   1.068   1.192

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.721250   0.072970   9.884 < 2e-16 ***
age -0.009953   0.001666 -5.974 2.32e-09 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27295 on 19997 degrees of freedom
Residual deviance: 27259 on 19996 degrees of freedom
AIC: 27263

Number of Fisher Scoring iterations: 4

Based on the results, the p value: < 2e-16 is less than the sign. Level(0.05) . Therefore we reject the null hypothesis and accept the H1, which confirms that age is a factor to contract the disease

B. Chi-Square Test
Ho = No Significant association between marital status and ethnic group
H1 = There is significant association between marital status and ethnic group

Pearson's Chi-squared test
data: project$ethnic and project$marital
X-squared = 6.4592, df = 5, p-value = 0.2641

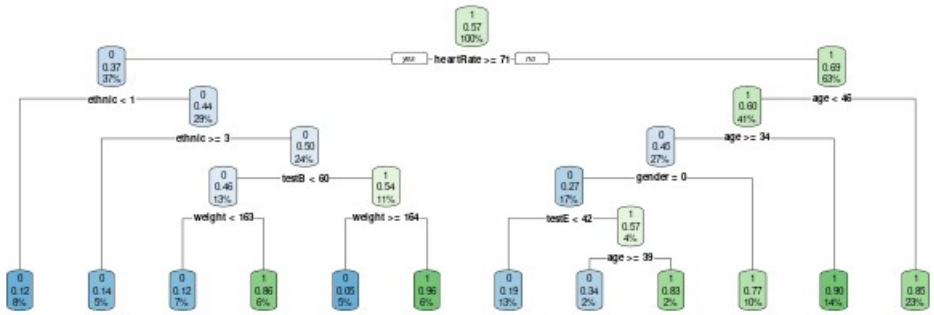
Based on the analysis, the p value (0.2641) is greater than the sign. level(0.05)
Therefore we accept the null hypothesis and reject the alternative hypothesis, which confirms that there is no significant association between marital status and ethnic group

C. Linear Regression
Ho = Age does not contribute to height
H1 = Age does contribute to height
Call:
lm(formula = age ~ height, data = project)
Residuals:
Min 1Q Median 3Q Max
-31.834 -5.840   1.119   6.166 33.180
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.280501   1.511981 28.625 <2e-16 ***
height -0.006772   0.023227 -0.292 0.771

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.608 on 19996 degrees of freedom
Multiple R-squared: 4.251e-06,	Adjusted R-squared: -4.576e-05
F-statistic: 0.085 on 1 and 19996 DF, p-value: 0.7706

Based on the analysis, the p value(0.7706) is greater than the sign. Value(0.05). Therefore we accept the null hypothesis, therefore Age does not contribute to height.

4. Decision Tree Visualization



A decision tree model was developed to further investigate the relationship between various factors and diseases in the patient data.
Model Performance:

The model seems to perform reasonably well overall, with an accuracy of 85.43%. This confusion matrix provides valuable insights into the model's performance and can help guide further refinement or decision-making based on the model's predictions.

6. Insights and Conclusions

Key findings from our analysis include:

1. There's a no significant association between marital status and ethnic group
2. The age group of about 30 years to about 50 years were over represented
3. Those without the disease is heavily concentrated in the younger age range.
4. The increasing variability in weight with age suggests that factors influencing weight become more diverse or impactful as people get older.
5. Age is a factor to contract the disease